Décompositions en éléments sonores et applications musicales

Mathieu Lagrange, Roland Badeau, Bertrand David, Nancy Bertin, Olivier Derrien, Sylvain Marchand, Laurent Daudet

Décompositions en Éléments Sonores et Applications Musicales

M. Lagrange[1], R. Badeau[2], B. David[2], N. Bertin[3], O. Derrien[4], S. Marchand[5], L. Daudet[6]

Ircam[1]*, CNRS STMS UPMC Paris, France*
Institut Télécom[2]*, Télécom ParisTech, CNRS LTCI Paris, France*
METISS Project[3]*, IRISA-INRIA Rennes, France*
LMA[4]*, CNRS - UPR 7051 Marseille, France*
LaBRI CNRS[5]*, Université de Bordeaux 1 Talence, France*
Université Paris Diderot[6]*, Institut Langevin CNRS UMR 7587 Paris, France*
Contact: `mathieu.lagrange@ircam.fr`

RÉSUMÉ. *Dans cet article sont présentés de manière synthétique les résultats du projet ANR DESAM (Décompositions en Éléments Sonores et Applications Musicales). Ce projet comportait deux parties, la première portant sur des avancées théoriques de techniques de décompositions de signaux audionumériques et la seconde traitant d'applications musicales de ces décompositions. La plupart des aspects abordés dans le projet ont donné lieu à de nouvelles méthodes et algorithmes qui sont regroupés au sein d'une boîte à outils, la DESAM Toolbox. Celle-ci rassemble un ensemble de fonctions Matlab® dédiées à l'estimation de modèles spectraux très utilisés pour les signaux musicaux. Les méthodes étudiées dans ce projet peuvent bien sûr être utiles pour la recherche automatique d'informations dans les signaux musicaux, mais elles constituent avant tout une collection d'outils récents pour décomposer les signaux selon différents modèles, avec pour résultat des représentations mi-niveau variées, pouvant être utiles dans d'autres domaines d'application.*

ABSTRACT. *In this paper is presented the DESAM project which was divided in two parts. The first one was devoted to the theoretical and experimental study of parametric and non-parametric techniques for decomposing audio signals into sound elements. The second part focused on some musical applications of these decompositions. Most aspects that have been considered in this project have led to the proposal of new methods which have been grouped together into the so-called DESAM Toolbox, a set of Matlab® functions dedicated to the estimation of widely used spectral models for music signals. Although those models can be used in Music Information Retrieval (MIR) tasks, the core functions of the toolbox do not focus on any specific application. It is rather aimed at providing a range of state-of-the-art signal process-*

e soumission à *Traitement du signal*, le 14 octobre 2011

ing tools that decompose music recordings according to different signal models, giving rise to different "mid-level" representations.

MOTS-CLÉS : *Traitement du signal audio, modèles spectraux, modélisation du son*

KEYWORDS: *Audio processing, Spectral models, Sound modeling*

1. Introduction

Le traitement du signal musical a effectué des progrès importants lors de la dernière décennie, tant d'un point qualitatif que quantitatif. Son domaine d'activité s'est maintenant élargi bien au-delà de son champ d'application traditionnel issu du monde de l'ingénierie sonore "audio" (composants "hi-fi" et traitement du son adapté, effets sonores, codage audio, ...) : il est maintenant en interaction forte avec d'autres champs disciplinaires tels que la perception de la musique et la psychoacoustique, la recherche d'information musicale ("Music Information Retrieval", ou simplement MIR), le design sonore, les outils à la composition et le "home-studio", les interfaces homme-machine ... pour ne citer que les plus répandus.

Quel que soit le point de vue adopté, il est couramment admis que l'un des objectifs à long terme de ces recherches serait de pouvoir décomposer un signal musical complexe, multi-instrumental, en entités élémentaires possédant chacune un "sens" intrinsèque, ce qu'il est possible d'appeler des "objets" musicaux (Schaeffer, 1977). Il est important de souligner que l'information portée par ces objets ne l'est pas seulement au sens du traitement du signal, mais aussi d'un point de vue musicologique ou perceptif (à l'opposé des décompositions "atomiques" temps-fréquence ou temps-échelle dans un contexte d'analyse/synthèse, ou des "caractéristiques" du signal dans un système de reconnaissance automatique de type MIR). Cette recherche d'objets élémentaires porteurs de "sens" est parfois appelée "analyse mi-niveau" dans la littérature, et elle est aussi liée au champ de recherche de l'Analyse Computationnelle de Scène Auditive (CASA) (Bregman, 1990). A partir de cette information, il doit être possible de déterminer les notes qui ont été jouées (tâche de transcription), avec quel type d'instrument, avec quelle technique de jeu ou encore les paramètres acoustiques des conditions d'enregistrement. Ainsi dans notre cas, les "objets" ou "éléments sonores" pourront être des notes, des groupes de notes (accords), ou des éléments structurels d'une note (par exemple ses harmoniques, ou son attaque).

Pour traiter ces tâches complexes, en tenant compte de la grande variété des musiques disponibles, des degrés de polyphonie multiples, des instrumentations diverses, les outils ne peuvent être que pluriels. Le point de vue adopté dans le projet DESAM est d'aborder ce problème sous l'angle de la modélisation, la spécification d'un modèle étant le premier pas sur la voie de sa caractérisation. Le but du projet DESAM (Décomposition en Éléments Sonores et Applications Musicales, financé par l'ANR), est de développer et fournir un éventail d'outils de traitement du signal récents, parfois encore méconnus, et complémentaires, qui permettent de décomposer des signaux musicaux réels (par exemple issus d'enregistrements du commerce) afin d'obtenir différents types de représentations "mi-niveau". En d'autres termes, les coefficients des modèles constitueront une nouvelle représentation du son. Les avancées de ce projet sont rassemblées et disponibles sous forme d'une boîte à outils, de manière à favoriser une utilisation la plus large possible, en particulier par la communauté académique dans le cadre du prototypage rapide de nouveaux algorithmes. Nous avons pris le parti

de la proposer comme un jeu de fonctions Matlab®, disponible en ligne[1], sous licence GPL. Elle est maintenue et améliorée en fonction des échanges avec la communauté d'utilisateurs de par le monde.

Dans cet article, après avoir présenté le cadre du projet DESAM, et motivé le développement de cette nouvelle boîte à outils dans la Section 2, nous décrirons les différents modèles de signaux étudiés lors du projet. Les outils de base, modèles sinusoïdaux ou plus généralement modèles spectraux, seront décrits dans les sections 3 et 4, respectivement. Enfin, nous présenterons en Section 5 une application à la transcription musicale automatique.

2. Le projet DESAM

Le projet DESAM [2] est un projet de recherche fondamentale qui a réuni quatre laboratoires français :

– CNRS LTCI (Laboratoire Traitement et Communication de l'Information), Institut Télécom - Télécom Paristech, Paris, France ;

– LAM (équipe Lutheries - Acoustique - Musique), Institut Jean Le Rond d'Alembert, UPMC, Université Paris 6 ;

– LaBRI (Laboratoire Bordelais de Recherche en Informatique), Université Bordeaux 1 ;

– STIC (Laboratoire Sciences et Technologies de l'Information et de la Communication), Université de Toulon et du Var.

Conduit par le LTCI, le projet a démarré en novembre 2006 et s'est poursuivi jusqu'en février 2010. Il comprenait deux parties. La première était dédiée à l'étude théorique et expérimentale de techniques paramétriques et non paramétriques pour décomposer des signaux audio en éléments sonores. La seconde portait sur des applications musicales de ces décompositions.

2.1. *Décompositions en éléments sonores*

Une caractérisation simplifiée des notes de musique consiste à préciser leur hauteur et leur timbre, celui-ci étant compris comme l'identification de la source (instrument). Le modèle sinusoïdal, qui représente ces notes comme un mélange de sinusoïdes et nécessite l'estimation correcte des fréquences et amplitudes correspondantes, ainsi que de leurs variations temporelles, est largement utilisé et pertinent pour analyser ces dimensions de hauteur et de timbre. Dans ce projet, nous avons développé des méthodes à haute résolution (HR), innovantes pour l'analyse temps-fréquence, dans le but d'estimer les variations temporelles fines des paramètres de

1. http ://www.tsi.telecom-paristech.fr/aao/2010/03/29/desam-toolbox
2. http ://perso.telecom-paristech.fr/rbadeau/desam

fréquence et d'amplitude (Badeau and David, 2008a; David *et al.*, 2006; David and Badeau, 2007). La modélisation et l'estimation de sinusoïdes non stationnaires a été approfondie dans (Marchand and Depalle, 2008).

Par ailleurs, un morceau de musique est composé d'éléments sonores (dont souvent de multiples notes), dont la combinaison temporelle et fréquentielle fournit le sens. Cette description à base d'agencement d'un nombre limité d'éléments sonores (qui peuvent être soit des notes isolées, des combinaisons de notes, ou des parties de notes), a conduit naturellement à l'utilisation d'une représentation appelée *parcimonieuse* (Daudet, 2010), ce nombre limité d'éléments sonores permettant de décrire tout le contenu musical. Une autre approche se base sur la factorisation en matrices positives (*non-negative matrix factorization*, (NMF)), capable d'exploiter les redondances d'un morceau de musique (note répétée plusieurs fois par exemple), pour inférer automatiquement les éléments sonores, chaque élément étant représenté par ses caractéristiques spectrales et ses occurrences au cours du temps. Ces techniques ont été raffinées pour mieux s'adapter aux propriétés des signaux musicaux étudiés (Bertin *et al.*, 2009b; Hennequin *et al.*, 2010).

2.2. *Applications musicales*

Analyser un enregistrement polyphonique afin d'en extraire ou de modifier son contenu musical (par exemple les instruments, le rythme ou les notes) est un exercice difficile, même pour un musicien expérimenté. La finalité du projet DESAM était de rendre une machine capable d'effectuer de telles tâches. En voici quelques unes :

– La capacité d'identifier des instruments de musique à partir d'un enregistrement est un atout pour l'indexation de musique. Une caractéristique importante d'un son qui définit la perception du timbre en est l'enveloppe spectrale.

– La capacité d'estimer la hauteur d'un son (sur une échelle allant du grave vers l'aigu) est cruciale pour identifier des notes de musique, mais reste difficile dans le cas d'un enregistrement polyphonique, à cause du recouvrement des sons.

– Si produire un son d'après une partition de musique est facile pour un musicien comme pour un ordinateur, le problème inverse, appelé *transcription automatique*, qui vise à retrouver la partition de musique à partir d'un enregistrement, s'avère bien plus difficile et requiert le savoir-faire d'un expert.

– Stocker et transmettre un volume toujours croissant d'enregistrements musicaux nécessite de coder ces données dans un format aussi compact que possible, en faisant un compromis entre la quantité d'information codée et la qualité du son reproduit.

Les décompositions en éléments sonores fournissent une représentation du signal comme une somme d'entités élémentaires. A partir de ces entités, des descripteurs de haut niveau sont extraits et sont utilisés pour la reconnaissance d'instruments de musique, l'estimation de rythme et l'estimation de hauteurs multiples. Ces tâches sont toutes requises dans la conception d'un transcripteur automatique.

Nous avons donc proposé de nouvelles méthodes pour estimer et comparer les enveloppes spectrales de sons musicaux (Badeau and David, 2008b; Lagrange *et al.*, 2010a). Nous avons aussi proposé des méthodes originales d'estimation de hauteurs, capables d'estimer jusqu'à 10 notes simultanées (Emiya *et al.*, 2010; Badeau *et al.*, 2009), qui ont été utilisées dans un algorithme de transcription automatique conçu pour la musique de piano (Emiya *et al.*, 2008). Une méthode alternative de transcription basée sur la NMF a également été développée pour une classe plus large d'instruments de musique (Bertin *et al.*, 2010a). Par ailleurs, la précision de la décomposition a permis de réaliser une analyse physique de la production de sons dans des instruments de musique (Le Carrou *et al.*, 2009), et le développement de méthodes plus efficaces pour coder et modifier les sons. Deux approches ont été retenues pour le codage. La première, basée sur des méthodes HR, a permis d'atteindre de très bas débits (Derrien *et al.*, 2008). La seconde, basée sur des décompositions parcimonieuses, a conduit à un codeur audio évolutif qui peut atteindre la transparence (Ravelli *et al.*, 2008). Des modifications sonores ont été réalisées, soit en ré-échantillonnant les paramètres d'un modèle sinusoïdal (Raspaud and Marchand, 2007), soit en modifiant les éléments sonores d'une décomposition parcimonieuse (Derrien, 2007).

2.3. *Boîte à outils DESAM*

A l'heure actuelle la majorité des outils existants sous Matlab sont centrés sur deux objectifs :

– soit ils sont liés à la perception, avec par exemple l'Auditory Toolbox (Slaney, 1998) et la Computer Audition Toolbox (CATbox) (Dubnov and Yazdani, 2010). Dans ce cas le but principal est de proposer des outils de pré-traitement pertinents d'un point de vue perceptif, avant une analyse plus poussée,

– soit ils ont été développés dans le contexte du Music Information Retrieval (MIR Toolbox (Lartillot and Toiviainen, 2007), MA Toolbox (Pampalk, 2004)), auquel cas ils sont basés sur des paramètres signal relativement rudimentaires, les plus souvent extraits après simple segmentation temporelle du signal.

Nous pensons qu'il existe dans la communauté un besoin d'outils qui permettent d'implémenter les outils les plus avancés en analyse des signaux, et qui soient capables non seulement d'analyser des enregistrements musicaux selon différents modèles de signaux - et donc utilisables comme paramètres d'entrée de systèmes MIR ; mais aussi de re-synthétiser la musique à partir des paramètres extraits - en d'autres termes, de pouvoir utiliser le système d'analyse/synthèse comme un système de compression audio au sens de l'ingénierie, ou comme un système de codage neuronal au sens de la communauté des sciences cognitives. Ces techniques récentes, comme les méthodes à haute-résolution adaptées à l'audio, ou la NMF, ont déjà fait la preuve de leur pertinence dans un grand nombre de cas pratiques. Il est maintenant utile de les proposer à une plus grande communauté d'utilisateurs afin qu'ils puissent devenir des outils "classiques", de la même façon que la transformée de Fourier à fenêtre glissante, ou la synthèse additive, ont été adoptées dans les dernières décennies. Les utilisateurs

pourront remarquer que le cœur de chacun de ces algorithmes est systématiquement très simple, particulièrement compact, et nous pensons qu'ils ne demandent qu'à être les fondations robustes d'une multitude de développements et d'améliorations.

Pour ce faire, la majeure partie des contributions développées dans le projet DESAM sont regroupés dans la boîte à outils DESAM, sous forme d'une compilation de fonctions écrites dans le langage Matlab et distribué sous licence "GNU General Public License" (GPL). La boîte à outils est composée de trois parties, matérialisées par différents répertoires. Les deux premières parties sont dédiées au cœur du projet : la représentation du signal sonore par des modèles sinusoïdaux ou spectraux. La dernière partie est dédiée à une application des ces modèles pour la transcription de la musique polyphonique.

3. Modèles sinusoïdaux

3.1. *Modélisation à court terme*

Étant donnée une petite fenêtre temporelle d'observation d'un signal audio, il est possible d'estimer les fréquences et les amplitudes des sinusoïdes qui le composent. La boîte à outils DESAM fournit plusieurs alternatives pour réaliser cette tâche. Une première approche consiste à estimer ces paramètres ainsi que leurs dérivées à partir du spectre (discret) de Fourier. Une seconde approche restreint un peu le modèle sinusoïdal mais permet de lever la contrainte de résolution fréquentielle de la transformée de Fourier discrète : il s'agit des méthodes à haute résolution (HR).

3.1.1. *Méthodes basées sur la transformée de Fourier*

La boîte à outils DESAM contient une estimation efficace des paramètres du modèle sinusoïdal non stationnaire étudiée dans (Marchand and Depalle, 2008), basée sur la transformée de Fourier et des extensions des méthodes de la réallocation spectrale (Auger and Flandrin, 1995) ou de la dérivée (Lagrange and Marchand, 2007).

Il a été prouvé que ces deux méthodes étaient équivalentes en théorie comme en pratique, et qu'elles produisaient des résultats quasi optimaux en terme de précision de l'estimation (pourvu que la résolution en fréquence soit suffisante pour isoler les pics spectraux correspondant aux sinusoïdes), voir (Marchand and Lagrange, 2006), (Marchand and Depalle, 2008), (Hamilton *et al.*, 2009).

Bien que la résolution soit limitée à la largeur d'un canal fréquentiel de la transformée de Fourier discrète, (voir la Figure 1 pour un exemple du phénomène d'interférence fréquentielle), la précision est, elle, quasi optimale. De plus, le modèle non stationnaire considéré est plus général que celui utilisé par l'analyse HR, puisqu'il considère une modulation de la fréquence à un ordre supérieur :

$$s(t) = \sum_{p=1}^{P} a_p(t) \exp(j\phi_p(t)),$$

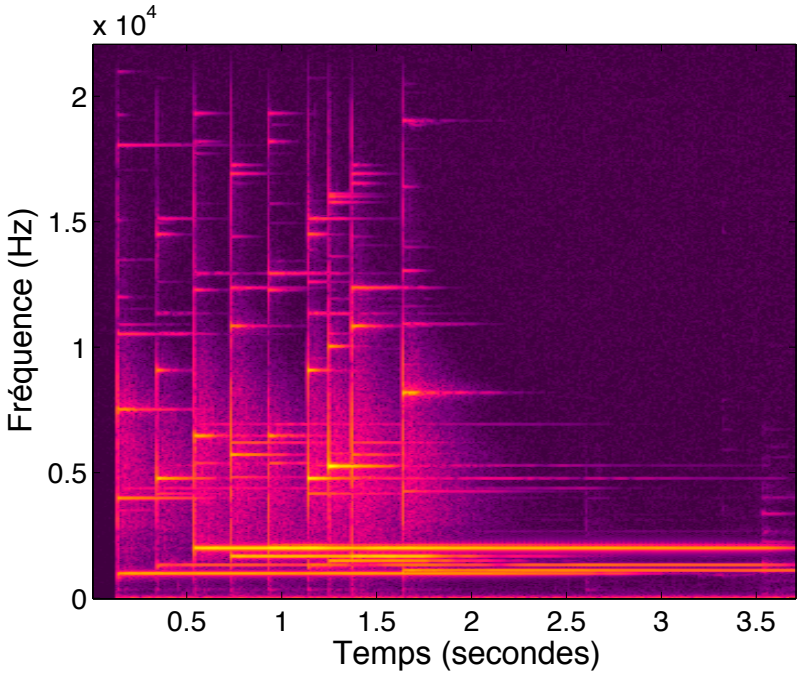

Figure 1 – *Spectrogramme de Fourier de notes de glockenspiel.*

où P est le nombre de partiels et

$$\begin{aligned} a_p(t) &= a_p \exp(\mu_p t), \\ \phi_p(t) &= \phi_p + \omega_p t + \tfrac{1}{2}\psi_p t^2. \end{aligned}$$

3.1.2. *Analyse/synthèse à Haute Résolution (HR) élémentaire*

Une méthode d'analyse / synthèse de signaux de musique a été proposée dans (Badeau, 2005). Elle est basée sur le modèle *Exponentially Damped Sinusoids* (EDS) :

$$s(t) = \sum_{p=1}^{P} \alpha_p \, z_p^t,$$

Ce modèle est estimé à l'aide d'une méthode HR sous-espace nommée ESPRIT, qui s'affranchit de la limite de résolution spectrale de la transformée de Fourier, et effectue une estimation précise des composantes sinusoïdales du signal. Pour commencer, le signal est pré-accentué. L'enveloppe du bruit auto-régressif est estimée, et le signal est blanchi avec un filtre à réponse impulsionnelle finie (Badeau, 2005).

3.1.3. *Analyse par sous-bande et segmentation dynamique pour le codage audio HR*

Pour une application efficace au codage audio, l'analyse/synthèse HR peut être réalisée dans les sous-bandes de fréquences et à partir d'une segmentation trame-à-trame dynamique. Pour réduire la complexité, on utilise un banc de filtres à reconstruction parfaite, qui conserve seulement les fréquences positives du signal. La segmentation trame-à-trame dynamique est effectuée à l'aide d'un algorithme de détection des attaques (Duxbury *et al.*, 2002). Le script correspondant est `demo.m`. Le codec audio décrit dans (Derrien *et al.*, 2008) est basé sur cette méthode d'analyse/synthèse.

3.1.4. *Analyse adaptative en sous-bande et poursuite rapide de sous-espace*

Une version adaptative de cette méthode, utilisant un banc de filtre classique (avec blanchiment sous-bande par sous-bande) et une analyse/synthèse à longueur de trame fixe, est également décrite dans (Badeau, 2005). Cette technique conduit à une nouvelle représentation, appelée HRogramme, où les composantes du signal sont représentées comme des points dans le plan temps-fréquence (cf. Figure 2). La partie stochastique est alors définie comme le résiduel de cette décomposition. Les parties déterministe et stochastique peuvent ensuite être traitées séparément, ce qui permet d'appliquer des effets sonores de haute qualité.

3.2. *Modèles à long terme*

Popularisés par les travaux de MacAulay et Quatieri (McAulay and Quatieri, August 1986) pour le traitement de la parole ainsi que ceux de Xavier Serra (Serra and Smith, 1990) pour le traitement des signaux musicaux, la plupart des algorithmes de suivi de partiels relient dans le temps des pics spectraux dont les paramètres sont estimés par des méthodes à court-terme comme celles présentées précédemment. Ces pics ainsi reliés forment alors des *partiels*, des oscillateurs sinusoïdaux dant les paramètres de contrôle comme la fréquence ou l'amplitude évoluent lentement avec le temps.

Ces méthodes, que nous nommerons respectivement `Maq` et `Serra` utilisent des heuristiques comme la proximité en fréquence entre pics de trames successives de manière à assurer la variation lente des paramètres. Des approches plus sophistiquées ont été proposées dans (Lagrange *et al.*, 2007), de manière à assurer cette contrainte d'évolution lente des paramètres. On considère tout d'abord que des évolutions sont prédictibles, et qu'une modélisation Auto-Régressive (AR) de l'évolution passée des paramètres permet de mieux guider la recherche de continuation. Cette méthode est appelée `LP` (pour *Linear Prediction*) et un exemple de suivi des harmoniques d'une note de clarinette grâce à cette méthode est présentée sur la Figure 3. Ensuite, on retient parmi les continuations possibles celle qui engendre une évolution des paramètres la plus lisse possible (spectres associés de plus basse fréquence possible). Cette méthode est appelée `HF` (pour détection des Hautes Fréquences).

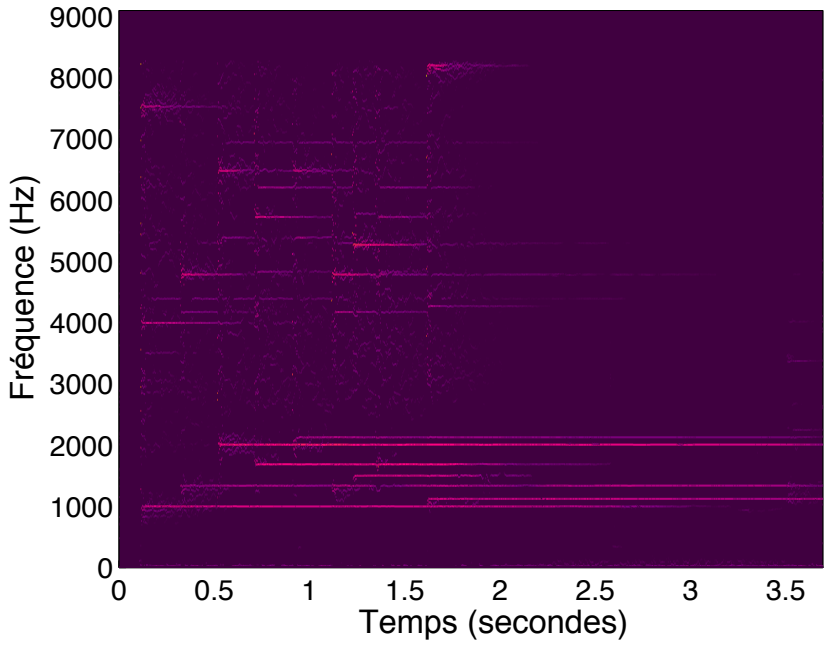

Figure 2 – *Estimation des fréquences de notes de glockenspiel utilisant la méthode rapide de poursuite de sous-espace. Le recouvrement fréquentiel est notablement réduit par rapport à un spectrogramme calculé à partir d'une transformée de Fourier classique, voir Figure 1.*

3.2.1. *Poursuite adaptative des composantes fréquentielles : l'algorithme HRHATRAC*

HRHATRAC signifie *High Resolution HArmonics TRACking* (poursuite haute résolution des harmoniques). Comme dans le paragraphe précédent, il s'agit d'un algorithme destiné à modéliser les sons comme somme de trajectoires spectrales, et d'un bruit environnant additif. HRHATRAC s'appuie sur l'efficacité d'un des plus récents algorithmes de poursuite adaptative de sous-espace (Badeau *et al.*, 2005) et sur une méthode de gradient pour mettre à jour les estimées des pôles du signal. Cela rend possible la mise à jour de chaque composante fréquentielle *individuellement*, d'un instant d'analyse au suivant. HRHATRACK conduit ainsi à une représentation du contenu sinusoïdal du signal en termes de trajectoires spectrales (composantes fréquentielles lentement variables).

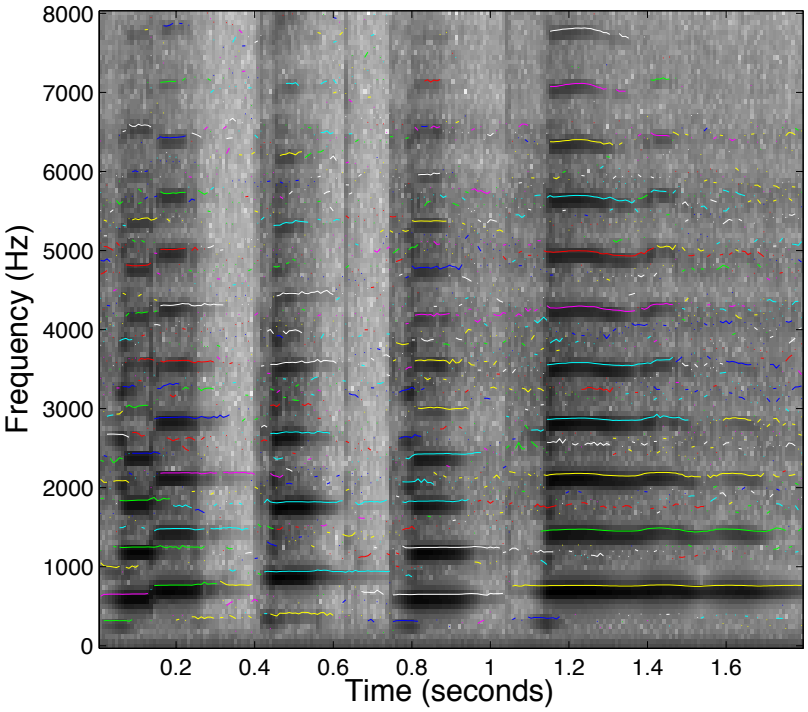

Figure 3 – *Résultats du suivi des harmoniques d'une note de clarinette grâce à la méthode LP.*

3.2.2. *Estimation adaptative des amplitudes instantanées*

Dans (David and Badeau, 2007) un schéma d'estimation adaptative est proposé pour mettre à jour les amplitudes instantanées à chaque échantillon. Ce schéma suppose connues les fréquences des composantes. Le résultat est une décomposition harmoniques+bruit, qui est calculée dans le même temps. Le cœur de cette analyse est un algorithme de moindre carrés rapide qui, connaissant une base de r composantes fréquentielles distinctes, met à jour récursivement les estimées des amplitudes et des phases. Alors qu'un calcul direct aurait nécessité $O(nr^2)$ opérations (n étant le nombre total d'échantillons traités), le coût principal de notre implémentation est seulement de $4r$ multiplications par échantillon.

4. Modèles spectraux et Applications

Si la modélisation du son sous forme de sinusoïdes est pertinente pour de nombreuses applications, il est également utile de considérer des représentations moins contraintes du contenu spectral d'un signal sonore.

4.1. *Représentations parcimonieuses*

Les représentations parcimonieuses ("sparse" (Plumbley *et al.*, 2010)), encore appelées décompositions atomiques ou représentations par dictionnaires de formes d'ondes, fournissent une approche complémentaire aux représentations paramétriques décrites précédemment, en ce sens que le modèle sous-jacent n'est pas nécessairement fixé *a priori*.

Ce domaine a connu des progrès remarquables dans la dernière décennie, tant au niveau théorique qu'algorithmique. En particulier, ces approches permettent une analyse des signaux à plusieurs niveaux : au niveau "atomique" pour une description au plus près du signal, puis au niveau de groupes d'atomes avec différents modèles de structures pour une description d'un plus haut niveau sémantique (Févotte *et al.*, 2008), (Daudet, 2006), (Moussallam *et al.*, 2011), (Ravelli *et al.*, 2010).

Ces outils ont été délibérément omis lors de la constitution de la boîte à outils DESAM car un certain nombre d'outils algorithmiques sont déjà disponibles, que ce soit sous Matlab pour un prototypage rapide (boîte à outils *sparsify*[3]), ou en C pour l'utilisation à des signaux musicaux de grande taille (plateforme MPTK[4]).

4.2. *Estimation d'enveloppes spectrales ARMA*

Nous avons proposé dans la référence (Badeau and David, 2008b) de nouveaux algorithmes pour estimer des modèles auto-régressifs (AR), à moyenne ajustée (MA) et ARMA dans le domaine spectral (alors que les méthodes classiques, comme la prédiction linéaire, travaillent généralement dans le domaine temporel). Ces algorithmes ont été obtenus par une approche de type maximum de vraisemblance, où des poids spectraux ont été introduits afin de pouvoir effectuer une estimation précise sur un ensemble de fréquences prédéfini, tout en ignorant les autres fréquences. Cela est particulièrement utile pour modéliser l'enveloppe spectrale de signaux harmoniques, dont le spectre ne contient qu'un ensemble réduit de coefficients pertinents (alors que la propriété d'harmonicité n'est pas prise en compte par les méthodes usuelles). Dans le cas simple d'un modèle AR, nous avons démontré que l'algorithme proposé converge vers la solution optimale, et que la vitesse de convergence est accélérée en déplaçant les pôles dans le domaine de stabilité à chaque itération. Dans le contexte du traite-

3. `http://users.fmrib.ox.ac.uk/~tblumens/sparsify/sparsify.html`
4. `http://mptk.irisa.fr/`

ment de la parole, nos résultats de simulation ont montré que la méthode proposée fournit une modélisation ARMA des voyelles nasales plus précise que la méthode de Durbin.

4.3. *Descripteurs de timbre*

La plupart des approches computationnelles traitant de la modélisation du timbre adoptent un encodage concis de l'enveloppe spectrale comme les coefficients cepstraux (MFCC pour "Mel-Frequency Cepstral Coefficients"). Nous avons étudié des descripteurs qui encodent l'évolution pseudo-périodique de certaines composantes du spectre au cours du temps. Nous avons montré dans (Lagrange *et al.*, 2010b) que ces descripteurs portent une information complémentaire aux descripteurs décrivant un spectre instantané comme les MFCC et que leur utilisation conjointe permet d'améliorer les capacités de discrimination de l'ensemble de description pour par exemple mieux discriminer les sons provenant de différents instruments de musique.

Ces descripteurs sont souvent utilisés pour déterminer la similarité entre une requête fournie par l'utilisateur et des éléments d'une base de données, ceci pour proposer à l'utilisateur des éléments de cette base qui sont "proches" au sens du timbre. On suppose généralement que les requêtes et les éléments de la base de données sont d'une qualité d'enregistrement équivalente. Or, ceci n'est pas vérifié dans un grand nombre de scénarios applicatifs où la requête est fortement dégradée. Dans ce cas, étudié dans (Lagrange *et al.*, 2010a), nous avons considéré une modélisation lisse de l'enveloppe spectrale pour les éléments de la base. Pour la requête, uniquement les pics proéminents sont retenus de par leur relative tolérance aux dégradations.

4.4. *Algorithme EM pour l'estimation de hauteurs multiples*

Le problème d'estimation de hauteurs multiples consiste à estimer les fréquences fondamentales de plusieurs sources harmoniques dont les partiels peuvent se superposer, à partir de leur mélange. Dans la référence (Badeau *et al.*, 2009), nous introduisons une nouvelle approche pour l'estimation de hauteurs multiples, s'inscrivant dans le cadre statistique de l'algorithme espérance-maximisation (EM), qui vise à maximiser la vraisemblance du spectre observé. La méthode proposée est particulièrement prometteuse, en raison de sa robustesse au recouvrement des partiels, et de sa capacité à simplifier la tâche d'estimation de hauteurs multiples en plusieurs estimations successives de hauteurs simples et d'enveloppes spectrales. Elle requiert une initialisation appropriée, incluant par exemple une première phase d'estimation élémentaire de hauteurs multiples, et pourrait avantageusement tirer parti d'heuristiques, afin d'éviter de rester bloqué dans des maxima locaux. L'efficacité de cette approche a été confirmée par nos simulations dans le contexte de l'identification d'accords de musique, effectuée sur des sons de synthèse.

5. Transcription automatique par factorisation du spectrogramme

Dans (Bertin *et al.*, 2010b), nous avons présenté une méthode originale pour la transcription automatique de musique polyphonique. Cette méthode est fondée sur un modèle bayésien incluant des contraintes d'harmonicité et de régularité temporelle à la NMF appliquée aux représentations temps-fréquence. Une variante algorithmique rapide, présentée dans (Bertin *et al.*, 2009a), est mise en œuvre dans la boîte à outils DESAM.

5.1. *Modèle*

Dans le modèle standard de la NMF pour la transcription musicale, une représentation temps-fréquence V à valeurs positives du signal (le module du spectrogramme, par exemple) est exprimée sous la forme d'un produit matriciel, illustré par un exemple simple sur la figure 4. Le facteur de gauche est un dictionnaire W de spectres de notes de musique, de taille fixée et inférieure aux dimensions du spectrogramme traité, et le facteur de droite H contient les activations temporelles correspondant à ces notes :

$$V \approx WH$$

Les facteurs sont appris simultanément et sans aucune supervision, par des règles de mise à jour multiplicatives faisant décroître un critère d'erreur.

5.2. *Mise en œuvre*

La représentation factorisée est une représentation temps-fréquence calculée sur une échelle non uniforme de fréquence, l'échelle "Equivalent Rectangular Bandwidth" (ERB). Les fréquences fondamentales des gabarits spectraux sont réglés régulièrement sur l'échelle MIDI accordée à 440 Hz. Le programme inclus dans la boîte à outils DESAM calcule ensuite la factorisation en minimisant une version pénalisée de la distance d'Itakura-Saito entre le spectrogramme original et sa reconstruction, les règles de minimisation étant dérivées du modèle bayésien du problème, sous leur forme rapide (Bertin *et al.*, 2009a). Le paramètre de forme de la loi inverse-Gamma contraignant la régularité des activations temporelles est réglé empiriquement, ainsi que le nombre de spectres appris, fixé à 88 (ceci étant le nombre de notes du piano). Une fois la factorisation effectuée, une détection par seuil est pratiquée sur les activations temporelles, la hauteur musicale de chacune y est associée et l'ensemble est converti dans un format textuel descriptif.

6. Discussion

L'estimation des paramètres de modèles sinusoïdaux stationnaires à partir de la transformée de Fourier est une méthode très largement utilisée pour modéliser des

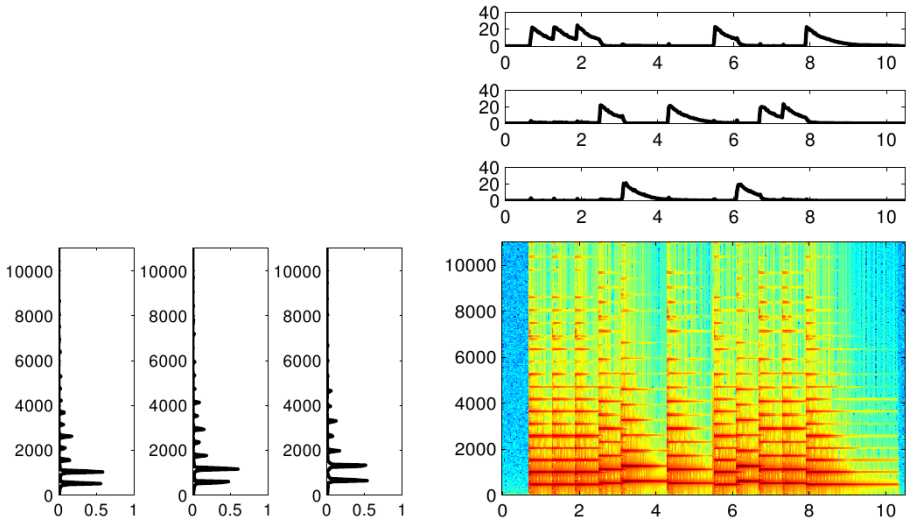

Figure 4 – Factorisation NMF d'une mélodie simple. En bas à gauche, les colonnes du dictionnaire de spectres W (avec l'axe des fréquences en vertical), en haut à droite les lignes d'activation temporelle H (avec l'axe temporel en horizontal), et en bas à droite, le spectrogramme V résultant du produit WH.

signaux musicaux. Dans ce cadre, nous avons étudié une classe particulière d'estimateurs améliorant la précision et l'extension de ces estimateurs pour les paramètres de second ordre.

L'utilisation de méthodes d'analyse spectrale à haute résolution était très rare dans le domaine de l'audio avant le commencement du projet, essentiellement en raison de leur forte complexité algorithmique. La mise au point d'algorithmes rapides et de prétraitements adéquats a permis leur utilisation dans un certain nombre d'applications courantes, dont certaines ont été abordées dans le cadre du projet (séparation, estimation du rythme, des hauteurs de notes, modélisation physique, codage audio). Aujourd'hui l'usage des méthodes à hautes résolution s'est répandu dans d'autres équipes de recherche en France et à l'étranger, notamment à l'université d'Aalborg au Danemark (Christensen *et al.*, 2008) et à l'université de McGill au Canada (Lagrange *et al.*, 2010c).

Les premières applications de la NMF en audio venaient de paraître lorsque le projet DESAM a démarré. Depuis, la NMF est devenu un outil très apprécié et très largement répandu en traitement des signaux de musique, notamment pour la transcription. Nos travaux ont permis de mieux adapter cet outil aux propriétés particulières de ces signaux (harmonicité, régularité spectrale, régularité temporelle, instationnarités).

En particulier, la NMF bayésienne rapide et contrainte présentée dans la section 5 s'est révélée plus performante que les approches NMF standard pour la transcription

automatique de sons réels de piano, et facilement extensible à d'autres instruments de musique. Toutefois, le modèle comprend certaines limitations intrinsèques. En particulier, il présuppose que le contenu spectral d'une note évolue peu (sous la forme d'un gain global) pendant son déroulement, ignorant des phénomènes comme le vibrato, ou l'extinction asynchrone des partiels d'une même note. De plus, la version contrainte est plus supervisée que la version standard (échelle de hauteurs musicales et gabarits spectraux partiellement fixés à l'avance) et nécessite en principe le choix manuel d'au moins trois paramètres (paramètre de forme pour la régularité temporelle, seuil de détection d'une note, paramètres réglant les gabarits spectraux)[5].

7. Conclusion

Nous avons présenté dans cet article l'ensemble des contributions du projet DESAM portant sur l'étude de nouveaux outils pour l'analyse des signaux sonores complexes que sont les enregistrements musicaux. Comme la plupart des modèles de signaux étudiés trouvent application dans de nombreux autres domaines du traitement du signal, certaines des avancées présentées devraient être pertinentes dans un cadre plus large que celui étudié lors de ce projet.

8. Bibliographie

Auger F., Flandrin P., « Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method », *IEEE Transactions on Signal Processing*, vol. 43, p. 1068-1089, 1995.

Badeau R., Méthodes à Haute résolution pour l'estimation et le suivi de sinusoïdes modulées. Application aux signaux musicaux., PhD thesis, École Nationale Supérieure des Télécommunications, Paris, France, April, 2005. Prix de thèse ParisTech 2006.

Badeau R., Boyer R., David B., « EDS parametric modeling and tracking of audio signals », *5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, p. 139-144, September, 2002.

Badeau R., David B., « Adaptive subspace methods for high resolution analysis of music signals », *Acoustics'08*, Paris, France, June-July, 2008a.

Badeau R., David B., « Weighted maximum likelihood autoregressive and moving average spectrum modeling », *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, Las Vegas, Nevada, USA, p. 3761-3764, March-April, 2008b.

Badeau R., David B., Richard G., « Fast Approximated Power Iteration Subspace Tracking », *IEEE Transactions on Signal Processing*, vol. 53, n° 8, p. 2931-2941, August, 2005.

Badeau R., David B., Richard G., « A new perturbation analysis for signal enumeration in rotational invariance techniques », *IEEE Transactions on Signal Processing*, vol. 54, n° 2, p. 450-458, February, 2006.

5. Par commodité d'utilisation, ces paramètres ont été fixés dans la version distribuée dans la boîte à outils DESAM.

Badeau R., Emiya V., David B., « Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra », *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, Taipei, Taiwan, p. 3073-3076, April, 2009.

Bertin N., Badeau R., Vincent E., « Fast bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription », *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, p. 29-32, 18-21 octobre, 2009a.

Bertin N., Badeau R., Vincent E., « Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, n° 3, p. 538-549, March, 2010a.

Bertin N., Badeau R., Vincent E., « Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, n° 3, p. 538-549, mars, 2010b.

Bertin N., Févotte C., Badeau R., « A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. », *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, Taipei, Taiwan, p. 1545-1548, April, 2009b.

Bregman A. S., *Auditory Scene Analysis : The Perceptual Organization of Sound*, The MIT Press, 1990.

Christensen M. G., Stoica P., Jakobsson A., Jensen S. H., « Multi-Pitch Estimation », *Signal Processing*, vol. 88, n° 4, p. 972-983, 2008.

Daudet L., « Sparse and structured decompositions of signals with the molecular matching pursuit », *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n° 5, p. 1808-1816, 2006.

Daudet L., « Audio sparse decompositions in parallel - Let the greed be shared ! », *IEEE Signal Processing Magazine, Special Issue on Signal Processing on Platforms with Multiple Cores : Part 2 – Design and Applications*, vol. 27, n° 2, p. 90-96, March, 2010.

David B., Badeau R., « Fast sequential LS estimation for sinusoidal modeling and decomposition of audio signals », *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, p. 211-214, October, 2007.

David B., Badeau R., Richard G., « HRHATRAC Algorithm for Spectral Line Tracking of Musical Signals », *International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, vol. III, Toulouse, France, p. 45-48, May, 2006.

David B., Richard G., Badeau R., « An EDS modelling tool for tracking and modifying musical signals », *Stockholm Music Acoustics Conference (SMAC 2003)*, vol. 2, Stockholm, Sweden, p. 715-718, August, 2003.

Derrien O., « Time-Scaling Of Audio Signals With Multi-Scale Gabor Analysis », *10th Conference on Digital Audio Effects (DAFX'07)*, Bordeaux, France, p. 1-6, September, 2007.

Derrien O., Richard G., Badeau R., « Damped sinusoids and subspace based approach for lossy audio coding », *Acoustics'08*, Paris, France, June-July, 2008.

Dubnov S., Yazdani M., « Computer Audition Toolbox (CATbox) », 2010. online web resource.

Duxbury C., Sandler M., Davies M., « A hybrid approach to musical note onset detection », *5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, September, 2002.

Ellis D. P. W., « PLP and RASTA (and MFCC, and inversion) in Matlab », 2005. online web resource.

Emiya V., Badeau R., David B., « Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches », *16th European Signal Processing Conference (EUSIPCO)*, Lausanne, Sweden, August, 2008.

Emiya V., Badeau R., David B., « Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, n° 6, p. 1643-1654, 2010.

Févotte C., Torrésani B., Daudet L., Godsill S., « Sparse linear regression with structured priors and application to denoising of musical audio », *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, n° 1, p. 174-185, 2008.

Hamilton B., Depalle P., Marchand S., « Theoretical and Practical Comparisons of the Reassignment Method and the Derivative Method for the Estimation of the Frequency Slope », *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, New Paltz, New York, USA, October, 2009.

Hennequin R., Badeau R., David B., « NMF with time-frequency activations to model nonstationary audio events », *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, Dallas, Texas, USA, p. 445-448, March, 2010.

Lagrange M., Badeau R., Richard G., « Robust similarity metrics between audio signals based on asymmetrical spectral envelope matching », *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, Dallas, Texas, USA, p. 405-408, March, 2010a.

Lagrange M., Marchand S., « Estimating the Instantaneous Frequency of Sinusoidal Components Using Phase-Based Methods », *Journal of the Audio Engineering Society*, vol. 55, n° 5, p. 385-399, May, 2007.

Lagrange M., Marchand S., Rault J., « Enhancing the Tracking of Partials for the Sinusoidal Modeling of Polyphonic Sounds », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, n° 5, p. 1625-1634, May, 2007.

Lagrange M., Raspaud M., Badeau R., Richard G., « Explicit Modeling of Temporal Dynamics within Musical Signals for Acoustical Unit Similarity », *Pattern Recognition Letters*, vol. 31, n° 12, p. 1498-1506, 2010b.

Lagrange M., Scavone G., Depalle P., « Analysis / Synthesis of Sounds Generated by Sustained Contact between Rigid Objects », *IEEE Transactions on Audio Speech and Language Processing*, vol. 18-3, p. 509-518, 2010c.

Lartillot O., Toiviainen P., « A Matlab toolbox for musical feature extraction from audio », *International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France*, 2007.

Le Carrou J.-L., Gautier F., Badeau R., « Sympathetic String Modes in the Concert Harp », *Acta Acustica united with Acustica*, vol. 95, n° 4, p. 744-752, July-August, 2009.

Marchand S., Depalle P., « Generalization of the derivative analysis method to non-stationary sinusoidal modeling », *11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, p. 281-288, September, 2008.

Marchand S., Lagrange M., « On the Equivalence of Phase-Based Methods for the Estimation of Instantaneous Frequency », *Proceedings of the 14th European Conference on Signal Processing (EUSIPCO'2006)*, EURASIP, Florence, Italy, September, 2006.

McAulay R. J., Quatieri T., « Speech analysis/Synthesis based on a sinusoidal representation », *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, n° 4, p. 744-754, August 1986.

Moussallam M., Daudet L., Richard G., « Audio signal representations for factorization in the sparse domain », *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2011)*, Prague, Czech Republic, May, 2011.

Pampalk E., « A Matlab Toolbox to Compute Similarity from Audio », *Proceedings of the ISMIR International Conference on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October, 2004.

Plumbley M., Blumensath T., Daudet L., Gribonval R., Davies M., « Sparse representations in audio and music : from coding to source separation », *Proceedings of the IEEE*, vol. 98, n° 6, p. 995-1005, 2010.

Raspaud M., Marchand S., « Enhanced resampling for sinusoidal modeling parameters », *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)*, New Paltz, New York, USA, p. 327-330, October, 2007.

Ravelli E., Richard G., Daudet L., « Union of MDCT bases for audio coding », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, n° 8, p. 1361-1372, November, 2008.

Ravelli E., Richard G., Daudet L., « Audio signal representations for indexing in the transform domain », *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n° 3, p. 434-446, 2010.

Schaeffer P., *Traité des objets musicaux*, Editions du Seuil, 1977.

Serra X., Smith J. O., « Spectral Modeling Synthesis : A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition », *Computer Music Journal*, vol. 14, n° 4, p. 12-24, 1990.

Slaney M., Auditory toolbox version 2, Technical report, Interval Research Corporation, 1998.

A. La boîte à outils DESAM

Cette annexe explicite pour chacune des méthodes décrites dans l'article, la fonction correspondante au sein de la boîte à outils DESAM qui est organisée selon une hiérarchie de répertoires correspondant à celle présentée dans la Figure 5.

A.1. *Modèles sinusoïdaux*

A.1.1. *Modélisation à court terme*

Méthodes basées sur la transformée de Fourier

Pour la méthode de la réallocation ("reassignment" en anglais), la syntaxe est de la forme :

```
[a, mu, phi, omega, psi, delta_t] =
      reassignment (x, Fs, m)
```

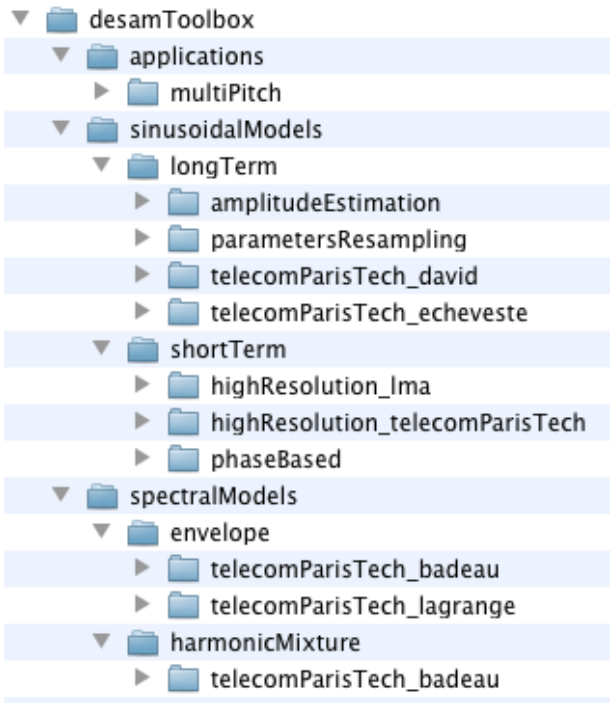

Figure 5 – *Hiérarchie de la boîte à outils DESAM*

Les paramètres en entrée sont :

– `x`, la trame de signal à analyser ;

– `Fs`, la fréquence d'échantillonnage (en Hz) ;

– `m`, l'indice du canal fréquentiel du pic spectral à considérer (optionnel) ;

et les valeurs en sortie sont `a`, `mu`, `phi`, `omega` et `psi`, correspondant respectivement aux valeurs estimées pour l'amplitude, la modulation d'amplitude, la phase, la fréquence et la modulation de fréquence du pic spectral. `delta_t` est le temps réalloué.

Pour la méthode de la dérivée, la syntaxe est similaire :

```
[a, mu, phi, omega, psi] =
derivative (x, d1, d2, Fs, m)
```

sauf que cette fonction nécessite les deux premières dérivées `d1` et `d2` du signal `x`, qui peuvent être calculées en utilisant la fonction

```
drv = discrete_derivative (src, Fs)
```

où `src` est le signal source and `drv` sa dérivée. Notons que la procédure de test de la dérivation discrète génère la Figure 1 de la référence (Marchand and Depalle, 2008). La procédure `test_global.m` génère les tests complets (Figures 2–6 de la référence (Marchand and Depalle, 2008)) avec des calculs lourds cependant, nécessitant donc du temps. Actuellement, l'implémentation de la méthode de réallocation est plus efficace. Par conséquent, nous recommandons cette méthode pour l'estimation des paramètres du modèle sinusoïdal. Référez-vous au script `test_example.m` dans le répertoire "sinusoidalModels/shortTerm/phaseBased" pour une démonstration mettant en œuvre cette méthode.

Analyse/synthèse à Haute Résolution (HR) élémentaire

La fonction `HR_analysis.m` dans le répertoire "sinusoidalModels/shortTerm/highResolution_lma" décompose le signal audio en une somme de sinusoïdes exponentiellement amorties sur chaque trame temporelle, et la composante de bruit est ignorée. La fonction `HR_synthesis.m` re-synthétise le signal audio correspondant à ces composantes EDS. Optionnellement, un filtre de blanchiment large-bande est proposé dans la fonction d'analyse. La syntaxe est :

```
[poles, amplitudes] = HR_analysis(signal, order, whitening)

signal = HR_synthesis(poles, amplitudes, signal_length)
```

Les paramètres sont :

- `signal`, signal à analyser et à synthétiser ;
- `order`, ordre du modèle ;
- `whitening`, drapeau pour la blanchiment optionnel ;
- `poles`, vecteur contenant les pôles du signal ;
- `amplitudes`, vecteur contenant les amplitudes complexes ;
- `signal_length`, longueur du signal à synthétiser (en échantillons).

Analyse par sous-bande et segmentation dynamique pour le codage audio HR

Le script correspondant est `demo.m`. Le codec audio décrit dans (Derrien *et al.*, 2008) est basé sur cette méthode d'analyse/synthèse.

Analyse adaptative en sous-bande et poursuite rapide de sous-espace

Le programme inclus dans la boîte à outils DESAM dans le répertoire "sinusoidalModels/shortTerm/highResolution_telecomParisTech" décompose le signal audio en une somme de sinusoïdes exponentiellement amorties et de bruit auto-régressif. Ensuite la méthode ESTER (Badeau *et al.*, 2006) est utilisée pour estimer le nombre de sinusoïdes, et un algorithme adaptatif rapide (appelé Itération Séquentielle) réalise la poursuite du sous-espace (Badeau *et al.*, 2002). La séparation sinusoïdes/bruit est

réalisée par projection sur l'espace signal et sur l'espace bruit (David *et al.*, 2003), et par reconstruction avec le banc de filtres de synthèse. L'ensemble des traitements est décrit dans la partie III de la référence (Badeau, 2005), et a été présenté lors de la conférence Acoustics'08 (Badeau and David, 2008a). La fonction principale est `analyse.m`, dont la syntaxe est la suivante :

```
[z,alpha,x] = analyse(s,Fs)
```

Les paramètres d'entrée et de sortie sont :

– `s`, signal à analyser ;

– `Fs`, fréquence d'échantillonnage (44100 Hz de préférence) ;

– `z`, matrice contenant les pôles du signal ;

– `alpha`, matrice contenant les amplitudes complexes ;

– `x`, partie sinusoïdale du signal.

Ce code a été appliqué avec succès à la décomposition de divers sons d'instruments de musique.

A.1.2. *Modèles à long terme*

Au sein de la boîte à outils, la syntaxe pour la méthode de suivi `Maq` est la suivante :

```
[P, Z,tag] = peaks2partialsMaq(A, F, tag, Z, deltaF)
```

où

– `A, F` sont respectivement les amplitudes et les fréquences des pics spectraux ;

– `tag` est le numéro de partiel assigné au prochain partiel à être créé ;

– `Z` contient l'état de chaque partiel actif ;

– `deltaF` détermine l'écart maximal entre les fréquences de deux pics consécutifs ;

– `P` contient les numéros assignés à chacun des partiels.

Les trois autres méthodes (`Serra`, `LP` et `HF`) ont une syntaxe similaire. Grâce à la notion d'état, cette syntaxe permet l'utilisation de ces méthodes en flux, comme présenté dans le script de démonstration `demo.m` dans le répertoire "sinusoidalModels/longTerm/telecomParisTech_echeveste". Ceci permet l'analyse de fichiers d'une durée arbitrairement grande.

On notera que ces quatre méthodes n'ont été validées que sur un nombre réduit d'exemples sonores. Ces méthodes sont donc fournies à titre de démonstration et ne doivent pas être considérées comme des implantations de référence, car pour cela, une validation plus solide est encore nécessaire.

Poursuite adaptative des composantes fréquentielles : l'algorithme HRHATRAC

La fonction DESAM correspondante, `hrhatrack`, renvoie pour la composante numéro p, sa fréquence instantanée $f_p(t) = (2\pi)^{-1}\phi'_p(t)$ et sa loi de modulation en amplitude instantanée $a_p(t)$. Le contenu sinusoïdal du signal (hors bruit) est alors obtenu comme $\sum_p a_p(t)\exp(j\phi_p(t))$.

La fonction principale est `hrhatrack.m` dans le répertoire "sinusoidalModels/longTerm/telecomParisTech_david", dont la syntaxe est la suivante :

```
[freqs,amps] = hrhatrack(s,Ns,P,beta,muL,muV)
```

Les paramètres d'entrée et de sortie sont :

– `s`, signal à analyser ;

– `Ns`, taille du sous-espace signal. Idéalement, le nombre de composantes fréquentielles s'il est connu. S'il ne l'est pas, il faut plutôt rechercher une surestimation de ce nombre par un facteur 1.5 à 2 ;

– `P`, taille de la matrice de covariance ($P \times P$), par défaut P = 3*Ns ;

– `beta`, facteur d'oubli pour la mise à jour de la matrice d'auto-covariance, par défaut : 0.99 ;

– `muL,muV`, pas du gradient pour la mise à jour respective des valeurs propres et des vecteurs propres, par défaut 0.9 ;

– `freqs,amps`, $f_p(t)$ et $a_p(t)$.

Un script de démonstration est fourni : `demo_hrhatrack.m`.

Estimation adaptative des amplitudes instantanées

La fonction principale est `fastls.m` dans le répertoire "sinusoidalModels/longTerm/amplitudeEstimation", dont la syntaxe est la suivante :

```
[b,xc,er] = fastls(s,z,n)
```

Les paramètres d'entrée et de sortie sont :

– `s`, signal à analyser ;

– `z`, z = [z1, z2, z3,...], pôles complexes ;

– `n`, longueur de la fenêtre d'analyse ;

– `b`, matrice des amplitudes instantanées, pour tout le signal ;

– `xc`, partie sinusoïdale du signal ;

– `er`, résiduel (`s = xc + er`).

Un script de démonstration est fourni : `demo1_fastls.m`.

A.2. *Modèles spectraux et Applications*

A.2.1. *Estimation d'enveloppes spectrales ARMA*

La fonction de démonstration principale est `test.m` dans le répertoire "spectralModels/envelope/telecomParisTech_badeau", qui effectue les simulations numériques présentées dans la référence (Badeau and David, 2008b), et affiche la figure correspondante.

A.2.2. *Descripteurs de timbre*

Le script `demo.m` dans le répertoire "spectralModels/envelope/telecomParisTech_lagrange" calcule un large ensemble de descripteurs spectro-temporels à partir d'un signal donné et affiche une matrice d'auto-similarité calculée à partir de ces descripteurs. Les descripteurs utilisant les MFCC se basent sur une implémentation Matlab publiée par Dan Ellis (Ellis, 2005).

A.2.3. *Algorithme EM pour l'estimation de hauteurs multiples*

La fonction principale est `test.m` dans le répertoire "spectralModels/harmonicMixture/telecomParisTech_badeau", qui réalise les simulations numériques présentées dans la référence (Badeau *et al.*, 2009), et affiche les figures correspondantes.

A.3. *Transcription automatique par factorisation du spectrogramme*

La fonction principale est `bertin_multipitch.m` dans le répertoire "applications/multiPitch/telecomParisTech_bertin" et s'utilise comme suit :

```
bertin_multipitch(wavfile,
framewise_f0file,notewise_f0file)
```

Les paramètres sont :

– `wavfile`, fichier audio d'entrée, à transcrire ;

– `framewise_f0file`, fichier de sortie contenant une transcription trame-à-trame (fréquence fondamentale f0 de chaque trame, toutes les 10ms) ;

– `notewise_f0file`, fichier de sortie contenant l'instant d'attaque, l'instant d'extinction et la fréquence fondamentale de chaque note.

B. Detailed abstract

Analyzing a polyphonic recording, in order to extract or to modify its musical content (e.g. the instruments, the beat, or the notes) is a difficult exercise, even for an

experimented musician. The tools described in this paper aim at making a computer able to perform such tasks. Let us mention three of them :

1) Pitch estimation. Estimating the pitch of a sound (on a scale from low to high) is critical for identifying musical notes, but remains difficult in a polyphonic recording, because of the overlap of sounds in the time and frequency domains.

2) Automatic transcription. If producing a sound given a musical score happens to be relatively easy both for the skilled musician and computer, the inverse problem, called "automatic transcription", which aims at recovering a musical score from a recording, proves to be much more complex and requires expert skills.

3) Audio coding. Storing and transmitting an increasing volume of musical recordings requires the coding of this data in a format that is as compact as possible. This involves a tradeoff between the quantity of coded information, and the quality of the reproduced sound.

In order to perform these tasks, one needs a model for polyphonic music. However, no single model can successfully account for all the characteristics of musical tones in general, and how they are intertwined with one another to form music. Musical notes are primarily characterized by their pitch and their timbre, specific to the instrument. They can thus be modeled as a mixture of sinusoids, whose frequencies and amplitudes are related to the pitch and timbre of the sound. In order to estimate the fine time variations of these two parameters, one needs precise analysis methods, such as the so-called "high-resolution" methods. Besides, since a musical piece is composed of multiple notes played at different times, it is naturally described as a combination of a number of elementary sound elements (which can be either isolated notes, combinations of notes, or parts of notes). Such a representation is called "sparse", since a very limited number of such sound elements, if well selected, should approximately describe the whole musical content. A complementary framework is based on a mathematical tool called "Non-negative Matrix Factorization" (NMF). It exploits the redundancies in a musical piece (a single tone being generally repeated within the piece), in order to identify the sound elements via their spectral characteristics and their various occurrences through time.

Amongst the results of the DESAM project, funded by the French ANR (Agence Nationale de la recherche), we have developed a number of analysis tools :

– an original pitch estimation method, capable of estimating up to ten simultaneous notes, which has been used in an automatic transcription algorithm for piano music.

– another transcription scheme based on NMF, which has been developed for a larger class of instruments.

– a coding method based on high-resolution analysis, which reaches very low bitrates (high compression ratio).

– another coding method, based on sparse decompositions, which is a scalable audio coder which can reach transparency (perceptively, the compressed sound cannot be distinguished from the original one).

Most aspects that have been considered in this project have led to the proposal of new algorithms which have been grouped together into the so-called DESAM Toolbox, a set of Matlab® functions dedicated to the estimation of widely used spectral models for music signals. This paper shortly presents the innovative tools that have been considered in order to build these systems. They are divided into two main parts : the first one is devoted to the theoretical and experimental study of parametric and non-parametric techniques for decomposing audio signals into sound elements ; the second part focuses on some musical applications of these decompositions.

Although these models can be used in a wide range of Music Information Retrieval (MIR) tasks, the core functions of the toolbox do not focus on any specific application. Their goal is rather to provide a wide range of state-of-the-art signal processing tools, that decompose music recordings according to different signal models, giving rise to different "mid-level" representations.

Mathieu Lagrange *a obtenu un doctorat en informatique en 2010 à l'université de Bordeaux 1. Il a poursuivi ses recherches portant sur la modélisation des signaux sonores durant deux post-doctorats respectivement a l'université de Victoria (BC, Canada) sous la direction de George Tzanetakis et à l'université McGill (QC, Canada) sous la direction de Philippe Depalle. Il a ensuite rejoint Telecom ParisTech pour participer aux projets Desam et Quaero. Il est actuellement chercheur CNRS à l'Ircam au sein de l'équipe analyse/synthèse des sons. Il est co-auteur de sept articles de revues, deux brevets, et d'une trentaine d'articles de conférences internationales à comité de lecture.*

Roland Badeau *a obtenu le diplôme d'ingénieur de l'École Polytechnique (X), Palaiseau, en 1999, puis le diplôme d'ingénieur de l'École Nationale Supérieure des Télécommunications (ENST), Paris, en 2001, le diplôme d'études approfondies (DEA) Mathématiques / Vision / Apprentissage de l'École Normale Supérieure (ENS), Cachan, en 2001, le doctorat de l'ENST en 2005 dans la spécialité traitement du signal et des images, et l'Habilitation à Diriger des Recherches (HDR) de l'Université Pierre et Marie Curie (UPMC), Paris, en 2010. Son doctorat a été récompensé par le Prix de thèse ParisTech 2006. En 2001, il a rejoint le Département de Traitement du Signal et des Images (TSI) de Télécom ParisTech (ENST), au sein du groupe Audio, Acoustique et Ondes (AAO), en tant que Chargé d'Enseignement et de Recherche, avant d'être promu Maître de Conférences en 2005. De novembre 2006 à février 2010, il a coordonné le projet ANR Jeunes Chercheuses, Jeunes Chercheurs intitulé "Décompositions en Éléments Sonores et Applications Musicales" (DESAM), qui rassemblait les efforts de quatre laboratoires français. Ses travaux de recherche portent, entre autres, sur le traitement des signaux audio appliqué à la musique, l'analyse spectrale à haute résolution et les décompositions non-négatives. Roland Badeau est Ingénieur en Chef du Corps des Mines, Senior Member de la société savante IEEE, et il est co-auteur d'une vingtaine d'articles de revues, de deux brevets, et d'une cinquantaine d'articles de conférences internationales à comité de lecture.*

Bertrand David *est né en 1967 à Paris, France. Il est diplômé de l'Université Paris-Sud Orsay et de l'Ecole Normale Supérieure de Cachan et il a obtenu l'Agrégation en 1991. Docteur*

en Acoustique et Traitement du Signal de l'Université Paris 6 depuis 1999, il a enseigné à l'ENSEA, une grande école dans le domaine du génie électrique, de l'électronique et de l'informatique et, dans le même temps, il a créé son entreprise de conseil en acoustique et traitement du signal, en travaillant en particulier sur des algorithmes embarqués de synthèse sonore pour mobiles. Depuis septembre 2001, il travaille comme Maître de Conférence à Telecom ParisTech, au sein du département Traitement du Signal et des Images où ses thématiques concernent plusieurs aspects du traitement des signaux audiofréquences tels que les modèles et représentations de ces signaux (il a coordonné une édition spéciale de la revue IEEE TASLP sur ce thème, qui a vu le jour en 2010), des applications comme la transcription automatique et l'utilisation d'algorithmes haute résolution en acoustique musicale. Membre du bureau du Groupe Spécialisé d'Acoustique Musicale depuis 10 ans, il a organisé des journées thématiques en lien avec ces thématiques et également participer depuis plusieurs années à une action destinée à établir des liens entre science et facture instrumentale.

Olivier Derrien *a obtenu le diplôme d'ingénieur de l'École Nationale Supérieure des Télécommunications (ENST), Paris, en 1998 puis le doctorat de l'ENST en 2002 dans la spécialité "Traitement du Signal et des Images". En 2003, il a été nommé Maître de Conférences à l'Institut d'Ingénieurs de Toulon et du Var (composante de l'Université du Toulon) et a rejoint le Laboratoire de Mécanique et d'Acoustique de Marseille dans l'équipe "Modélisation, Synthèse et Contrôle des Signaux Sonores et Musicaux" en 2008. Ses travaux recherche portent sur le traitement et la modélisation des signaux audio pour le codage compressif, les effets audionumériques et la synthèse de sons musicaux et d'environnement. Il est co-auteur de trois articles de revue et d'une quinzaine d'articles de conférences à comité de lecture.*

Sylvain Marchand *a obtenu son doctorat en 2000 dans le domaine de l'informatique musicale, récompensé par un accessit au prix de thèse SPECIF 2001. En 2001, il a été nommé Maître de Conférences au sein de l'équipe Image et Son du Laboratoire Bordelais de Recherche en Informatique (LaBRI), Université Bordeaux 1. En 2008, il a obtenu son Habilitation à Diriger des Recherches (HDR). Sylvain Marchand est Professeur à l'Université de Bretagne Occidentale depuis septembre 2011. Il y dirige la formation Image et Son à Brest. Sylvain Marchand est particulièrement actif dans les domaines de l'analyse, la transformation et la synthèse du son. Il est co-auteur de deux livres, une dizaine d'articles de revues, de deux brevets, et d'une cinquantaine d'articles de conférences internationales. Il est fortement impliqué dans le cycle de conférences DAFx (International Conference on Digital Audio Effects). Il est également éditeur associé des IEEE Transactions on Audio, Speech, and Language Processing depuis 2007.*

Nancy Bertin *a obtenu le diplôme d'ingénieur (2004) puis le diplôme de docteur en traitement du signal (2009) de Télécom ParisTech (ex-ENST). Elle est également titulaire du Master en Sciences et Technologies mention "Acoustique, Traitement du Signal et Informatique Appliqués à la Musique (Université Pierre et Marie Curie Paris 6 / IRCAM) obtenu en 2005. Son doctorat, consacré aux factorisations en matrices non-négatives et leur application à la transcription musicale, a été récompensé par le prix de thèse Signal, Image et Vision du GdR ISIS - Club EEA 2009.*

En 2010, elle a rejoint l'équipe-projet METISS de l'Inria (Centre Inria Rennes Bretagne Atlantique), en tant que chercheuse post-doctorante. Elle est désormais chargée de recherche

à l'IRISA (CNRS UMR 6074), dans la même équipe, depuis octobre 2011. Ses travaux actuels portent notamment sur les aspects théoriques et pratiques des représentations parcimonieuses structurées et de l'échantillonnage compressif de champs acoustiques.

***Laurent Daudet** a effectué des études de physique à l'Ecole Normale Supérieure et reçu en 2000 un doctorat en physique mathématique et modélisation de l'Université de Provence. Après un post-doctorat au Centre for Digital Music de Queen Mary University of London, UK, il est recruté en 2002 comme Maître de Conférences à l'UPMC, pour effectué des recherches au sein du Laboratoire d'Acoustique Musicale - maintenant intégré à l'Institut D'Alembert. Depuis 2009, il est professeur au département de physique à l'Université Paris Diderot, et rattaché à l'Institut Langevin "ondes et images". Depuis octobre 2010, il est également membre junior de l'Institut Universitaire de France. Laurent Daudet est Editeur Associé pour les IEEE Transactions on Audio, Speech and Language Processing, et auteur ou co-auteur de plus de 100 publications sur divers aspects de traitement du signal, principalement sur l'application des décompositions parcimonieuses à l'audio et l'acoustique.*